NoSQL Proof-of-Concept Proposal

Team name: PSG.LGD

Yichun Yan
Ziwei Jiang
Yifan Li
Weiqi Wang

September 16, 2019

Contents

1 Dota2 Game Replay Analysis

1.1 Potential Datasets

Our team aims at exploring data about a popular and long-lived computer game, Dota2. We will primarily get our data from the following three origins:

1. Dota2's official API
2. OpenDota API
3. Valve's Dota2 replay servers

We will get the match results data from Dota2's official API. Also we get the data we need to retrieve a replay from replay servers, then get replays. So basically we will have two different kinds of data for our further analysis, relatively large-scale but coarse-grained match results data, and very fine-grained replays data. Considering the file size of one replay, we decide to only include recent professional games for the replay data. Specifically, we can collect them by the following instructions:

1. Match Results:
   - Use GetMatchHistoryBySequenceNum API to get the match ids, we will need a a field called `start_at_match_seq_num` to specify the starting match sequence number of the results.
   - Then we can extract the last sequence number of the results as the `start_at_match_seq_num` for the next call. By doing this iteratively we can enlarge our dataset for our first kind of data.
2. Replays:
   - Get the match ids of recent professional games from the first dataset we collect.
   - Use the those ids via OpenDota API to get the information we need (`cluster` and `replay_salt`) for retrieving replays from Valve's replay server
   - Constructing links in this format: `http://replay<cluster>.valve.net/570/<match_id>_<replay_salt>.dem.bz2` to get the replays.

1.2 Datasets Description

We found the approaches to get those data on a developer's forum of Dota2. Valve's official document of many APIs is outdated so the forum is the only way for us to understand the resulting data.

The match results data is in JSON format, which contains snapshot information about a game's end, like how much gold one player earned at the end of the match, and also contains some information after a match, like how many players thumbs-up the game.

The JSON file for one match will be between 3KB~5KB, we planed to collect 1,000,000 matches' results data, which will add up to about 5GB.

The replay data is originally in binary format, but we found an open-source program to parse it into strings. So we will be handling the logs of all the activity happens in a single game. We may further parse this file considering that the file size of one single match and it contains many irrelevant information. The size of one replay's size is between 20MB~80MB. We plan to collect 1,000 matches' replays, which will add up to about 50GB.

1.3 Data Preprocessing and Specification

As the raw data is completely unstructured `.dem.bz2` file, preliminary preprocessing must be done before we store the data in our database.

Firstly, we can decompress the `.bz2` file with `bzip2` , which will give us a `.dem` file:

```
$ bzip2 -d data.dem.bz2
```

Next, we can utilize clarity, an open source Dota2 replay parser, to extract useful information from the `.dem` file. Based on our exploration and the examples provided by clarity, the following data will be available:

- Player name, id, team formation and hero choice
- Detailed log of the game, including a hero:
  - deals damage to another one
  - heals another one
  - receives/loses a buff/debuff
  - kills another one
  - uses his ability
  - uses an item
  - buys an item
  - receives/loses some gold
  - gains some XP
  - buys back (spending money in order to instantly re-spawn)
- Spawn/death of heros and NPCs
- Summary of each player's performance in the whole match, including:
  - Final level
  - Kills
  - Deaths
  - Assistance
  - Gold
  - Last hit
  - Deny

The following fields are critical to answering our Business Question, "???". They will need to be cleansed and validated.